doi:10.1093/nar/gkl858

PATRIC: The VBI PathoSystems Resource Integration Center

E. E. Snyder[1,*], N. Kampanya[1], J. Lu[1], E. K. Nordberg[1], H. R. Karur[1], M. Shukla[1], J. Soneja[1], Y. Tian[1], T. Xue[1], H. Yoo[1], F. Zhang[1], C. Dharmanolla[1], N. V. Dongre[1], J. J. Gillespie[1,4], J. Hamelius[1], M. Hance[1], K. I. Huntington[1], D. Jukneliene[2], J. Koziski[1], L. Mackasmiel[1], S. P. Mane[1], V. Nguyen[1], A. Purkayastha[1], J. Shallom[1], G. Yu[1], Y. Guo[1], J. Gabbard[1,3], D. Hix[1,3], A. F. Azad[4], S. C. Baker[2], S. M. Boyle[5], Y. Khudyakov[6], X. J. Meng[5], C. Rupprecht[6], J. Vinje[6], O. R. Crasta[1], M. J. Czar[1], A. Dickerman[1], J. D. Eckart[1], R. Kenyon[1], R. Will[1], J. C. Setubal[1] and B. W. S. Sobral[1]

[1]Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA, [2]Department of Microbiology and Immunology, Loyola University Medical Center, Maywood, IL 60153, USA, [3]Center for Human–Computer Interaction, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA, [4]Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD 21201, USA, [5]VA-MD Regional College of Veterinary Medicine, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA and [6]Centers for Disease Control and Prevention, Atlanta, GA 30333, USA

Received August 15, 2006; Revised September 14, 2006; Accepted October 1, 2006

ABSTRACT

The PathoSystems Resource Integration Center (PATRIC) is one of eight Bioinformatics Resource Centers (BRCs) funded by the National Institute of Allergy and Infection Diseases (NIAID) to create a data and analysis resource for selected NIAID priority pathogens, specifically proteobacteria of the genera *Brucella*, *Rickettsia* and *Coxiella*, and corona-, calici- and lyssaviruses and viruses associated with hepatitis A and E. The goal of the project is to provide a comprehensive bioinformatics resource for these pathogens, including consistently annotated genome, proteome and metabolic pathway data to facilitate research into counter-measures, including drugs, vaccines and diagnostics. The project's curation strategy has three prongs: 'breadth first' beginning with whole-genome and proteome curation using standardized protocols, a 'targeted' approach addressing the specific needs of researchers and an integrative strategy to leverage high-throughput experimental data (e.g. microarrays, proteomics) and literature. The PATRIC infrastructure consists of a relational database, analytical pipelines and a website which supports browsing, querying, data visualization and the ability to download raw and curated data in standard formats. At present, the site warehouses complete sequences for 17 bacterial and 332 viral genomes. The PATRIC website (https://patric.vbi.vt.edu) will continually grow with the addition of data, analysis and functionality over the course of the project.

INTRODUCTION

Bioterrorism became an important national security issue (1) following the deliberate release of anthrax spores into the US postal system in October 2001 (2). Meanwhile, emerging and reemerging infectious diseases (3) have had profound effects on public health in many parts of the world. Recognizing the pathogens responsible for these diseases as threats to homeland security, the National Institute of Allergy and Infectious Diseases (NIAID) of the US National Institutes of Health has embarked upon a series of initiatives aimed at developing a comprehensive understanding of the organisms identified as NIAID category A, B and C priority pathogens (for a complete list, see http://www3.niaid.nih.gov/biodefense/bandc_priority.htm). The Virginia Bioinformatics Institute's PathoSystems Resource Integration Center (PATRIC) is one of eight Bioinformatics Resource Centers (BRCs) established to

*To whom correspondence should be addressed. Tel: +1 540 231 5428; Fax: +1 540 231 2606; Email: eesnyder@vbi.vt.edu
Present addresses:
Y. Guo, NIAID Office of Technology Information Systems (OTIS), Contractor, MSD Inc., Bethesda, MD 20892, USA
J. D. Eckart, Advanced Technology Division, Aflac, Columbus, GA 31999, USA

study the NIAID priority pathogens and develop these information resources for the research community. While database resources for bacterial ((4) and those cited in (5)) and viral (6,7) genomics have been available for number of years, this project seeks to integrate genomics with comparative genomics and pathway analysis and ultimately proteomics, transcriptomics, immune epitope mapping, host-response and other downstream technologies. The goal is to help researchers and clinicians better detect and respond to biothreat agents (and infectious diseases in general) by facilitating the development of diagnostics, vaccines and therapeutics. This requires access to comprehensive information on the molecular biology, physiology and pathogenicity of these organisms.

THE PATHOSYSTEMS RESOURCE INTEGRATION CENTER

PATRIC is responsible for the eight organism categories listed in Table 1. The three genera of proteobacteria are all intracellular pathogens that are known or potential biowarfare agents. In the 1950s, *Brucella suis* was the first infectious agent developed for use as a biowarfare agent by the United States. Brucellosis, caused by *Brucella* sp., is an important agricultural disease infecting cattle, sheep, goats and swine as well as humans. It is highly contagious and readily dispersed as an aerosol (8). *Coxiella burnetii*, the causative agent of Q fever, is a highly infectious agent of relatively low lethality. Its interest as a biowarfare agent stems from its high infectivity, stability to heat and desiccation and potential for aerosol dispersal. The genus *Rickettsia* contains the organisms responsible for numerous types of typhus and arthropod-borne spotted fevers (9,10). *Rickettsia prowazekii* was developed as a bioweapon by the USSR in the 1930s and was used by the Japanese in Manchuria during World War II (11).

The five categories of viruses studied by PATRIC are all positive-strand ssRNA viruses, with the exception of Lyssaviruses, which have negative-strand ssRNA genomes. While there are no reports of any of these viruses being weaponized, they represent the causative agents for a number of emerging and reemerging diseases including Severe Acute Respiratory Syndrome (SARS), rabies and transmissible gastroenteritis. Recombinant vaccines for these viruses are either still in development or unavailable in areas where these infections are endemic or epidemic, compounding the public health risk.

The pace of research on these organisms has increased significantly since the turn of the millennium, with outbreaks, such as that of SARS in 2003 (12,13), spawning a flurry of scientific activity. The widespread use of automated DNA sequencing, microarray gene expression analysis and other high-throughput laboratory technologies has increased the volume of data produced, but not necessarily its accessibility. Currently, significant genomics and bioinformatics expertise is required to extract, process and interpret this wealth of data.

To address these problems, PATRIC has created an interdisciplinary team of bioinformaticians, software engineers, computational biologists and organism experts to build a publicly accessible resource aimed at providing high quality, analyzed and curated data to the infectious disease community working on these pathogens. To date, we have achieved the following objectives:

(i) collection and organization of existing genomic data for the eight pathosystems under a single, unified framework
(ii) genome annotation and curation following standardized procedures
(iii) visualization of raw data from analytical programs, as well as curated data
(iv) creation of orthologous gene groups within each organism category allowing comparative analysis of gene content
(v) prediction and visualization of bacterial metabolic pathways to complement functional analysis of proteins
(vi) integration of online literature reviews from PathInfo (14) for selected organisms.

Longer-term goals include integration of data from gene expression and proteomics experiments (including host-response), predicted protein and RNA secondary and tertiary structures, and well-cataloged literature compilations. Ultimately, we hope our website will become an essential tool for researchers working on these pathogens and provide networking opportunities within the pathogen research communities.

DATABASE DESCRIPTION

PATRIC is implemented on Oracle 9i RDBMS using the Genomics Unified Schema (GUS) version 3.5, developed at the Computational Biology and Informatics Laboratory at the University of Pennsylvania (see http://www.gusdb.org). GUS is used to store all sequence data and associated annotation with the exception of metabolic pathway data, which is

Table 1. PATRIC organisms

Organism category	Taxonomic rank	Organisms represented	Complete genomes	Approximate CDS count	Genome length	Related human diseases
Brucella	Genus	4	4	3150	3.3 Mb	Brucellosis
Coxiella	Genus	4	1	2002	2.0 Mb	Q Fever
Rickettsia	Genus	9	9	850–1550	1.2 Mb	Typhus, Rocky mountain spotted fever
Caliciviridae	Family	71	70	3	7.7 kb	Food-borne gastroenteritis
Coronavirus	Genus	184	169	7–12	29 kb	Severe acute respiratory syndrome (SARS)
Hepatitis A virus	Species	16	15	1	7.4 kb	Hepatitis
Hepatitis E virus	Species	48	48	3	7.2 kb	Hepatitis
Lyssavirus	Genus	13	12	5	12 kb	Rabies

Genome and organism counts reflect database content as of November 1, 2006.

stored in a separate schema developed for use with the Pathway Tools software (15) (see section 4.3). To minimize modifications to GUS and ensure compatibility with future releases, the Curation/Annotation Schema was developed to support PATRIC-specific curation activities.

The database is populated with all known full-length or nearly full-length genomic sequences for the eight organism categories listed in Table 1. Automated scripts query GenBank (16) daily to identify new or updated records. The corresponding sequences, annotation and associated literature are retrieved from NCBI and loaded following curatorial review to remove redundancies and assign unique names to each genome. RefSeq (17) records are used when available to take advantage of their more thorough and consistent annotation. Draft genome sequences from Joint Genome Institute (JGI)/Los Alamos National Labs (LANL) and the NIAID-funded Microbial Sequencing Centers will also be part of the PATRIC dataset. In addition to genome sequences and primary annotation from the original GenBank or RefSeq entry, the database stores the results of all automated and manual analyses described in the following section.

DATA ANALYSIS AND CURATION

Our motivation to invest resources in sequence-level annotation is to maintain a high standard of quality over time. Even when good reference annotation is available, there are many reasons to re-annotate microbial genomes (18). GenBank data are of variable quality and there is a trend towards depositing draft genome sequences with no annotation at all. In-house annotation also allows us to present supporting evidence and keep the annotation up to date. This is of particular importance for alignment-based annotation since databases such as GenBank (16) and UniProt (19) continue to grow at a prodigious rate.

Genome curation

Due to the large number of closely related genomes in each organism category, we have adopted an annotation strategy in which automated methods are applied to all genomes while detailed manual curation is applied to a limited number of reference genomes. The species *B.suis* 1330, *C.burnetii* RSA 493 and *R.prowazekii* str. Madrid E were chosen as reference genomes for their respective categories. Each viral category has (or will have) multiple reference genomes, representing phylogenetically diverse strains.

Automated nucleic acid and protein sequence annotation is accomplished using a Java-based genome annotation pipeline (unpublished), which reads an XML script containing the names and parameters of the analytical applications. The bacterial pipeline executes the gene prediction programs Glimmer (20) and GeneMark (21,22) followed by start site correction programs RBSfinder (23) and TICO (24). BLASTX (25) searches the non-redundant protein database, complementing the *ab initio* gene prediction methods. RNA genes are identified by tRNAscan-SE (26) and BLASTN searching against a ribosomal RNA database (27,28). The annotation protocol containing the full list of applications and parameters is available online at https://patric.vbi.vt.edu/documents/ under 'standard operating procedures'.

Results of the genome analysis pipeline are merged with original GenBank or RefSeq features for automated interpretation. A decision tree is used to classify genes into categories based on the level of agreement between the various prediction methods. Genes that are unambiguously predicted by multiple methods are automatically 'finalized', creating new 'gene', 'CDS' and/or '[t/r]RNA' features. The remaining genes are marked for manual curation. For viral genomes, an abbreviated pipeline is executed that emphasizes sequence alignment for gene identification and employs GeneMarkHMM optimized for mammalian (host) genomes.

Proteome curation

After curatorial review, finalized protein-coding (CDS) features are translated and subjected to another pipeline executing InterProScan and structure prediction methods such as MEMSAT 2 (29). Currently, each protein is associated with GO terms (30), TIGRroles, Enzyme Commission numbers (31) based on Pfam (32) and TIGRfam alignments (for a description of TIGRfam and TIGRroles, see: http://www.tigr.org/TIGRFAMs/). The protocol for automated proteome annotation is also available online. Manually curated protein sequences will be available in early 2007.

Once protein sequences are inferred from each genome in an organism category, putative ortholog groups are generated using BLASTP for all pairwise genome combinations and applying the conventional bidirectional-best-hit (BBH) criterion (33). While putative ortholog groups within the bacterial categories are generally well defined, many viral proteins cannot be readily clustered using the stringent BBH criterion. This is an active area of curation. Using the ortholog groups as a starting point, a reference protein list is created for each bacterial category consisting of the proteins of the reference genome (each representing one ortholog group) plus a representative protein from each ortholog group identified in the associated genomes. A gene occurring in only a single genome constitutes a 'group' of one and would be included in the reference list. The reference protein lists will be manually curated and include, whenever possible, detailed functional descriptions, gene symbols, GO terms and EC numbers. Thus, every protein in the database will either be manually curated or be linked to an ortholog group member that has been manually curated.

The ortholog groups are further processed to create multiple sequence alignments (MSAs) using MUSCLE (34) with default parameters. Phylogenetic estimations using the neighbor-joining method (35) were created based on trimmed alignments using PHYLIP (36). Trees were validated by bootstrapping (37) using a minimum of 100 replicates.

Metabolic pathway curation

To help users understand the function of the bacterial proteins in context, we have adopted the Pathway Tools system (15) to derive pathways from genome annotation and to fill potential gaps in annotation known as pathway holes. The system takes a list of protein names, descriptions and EC numbers as input. Proteins with EC numbers can be assigned roles directly; the roles of other proteins are suggested by lexicographic analysis of descriptive information and/or analysis

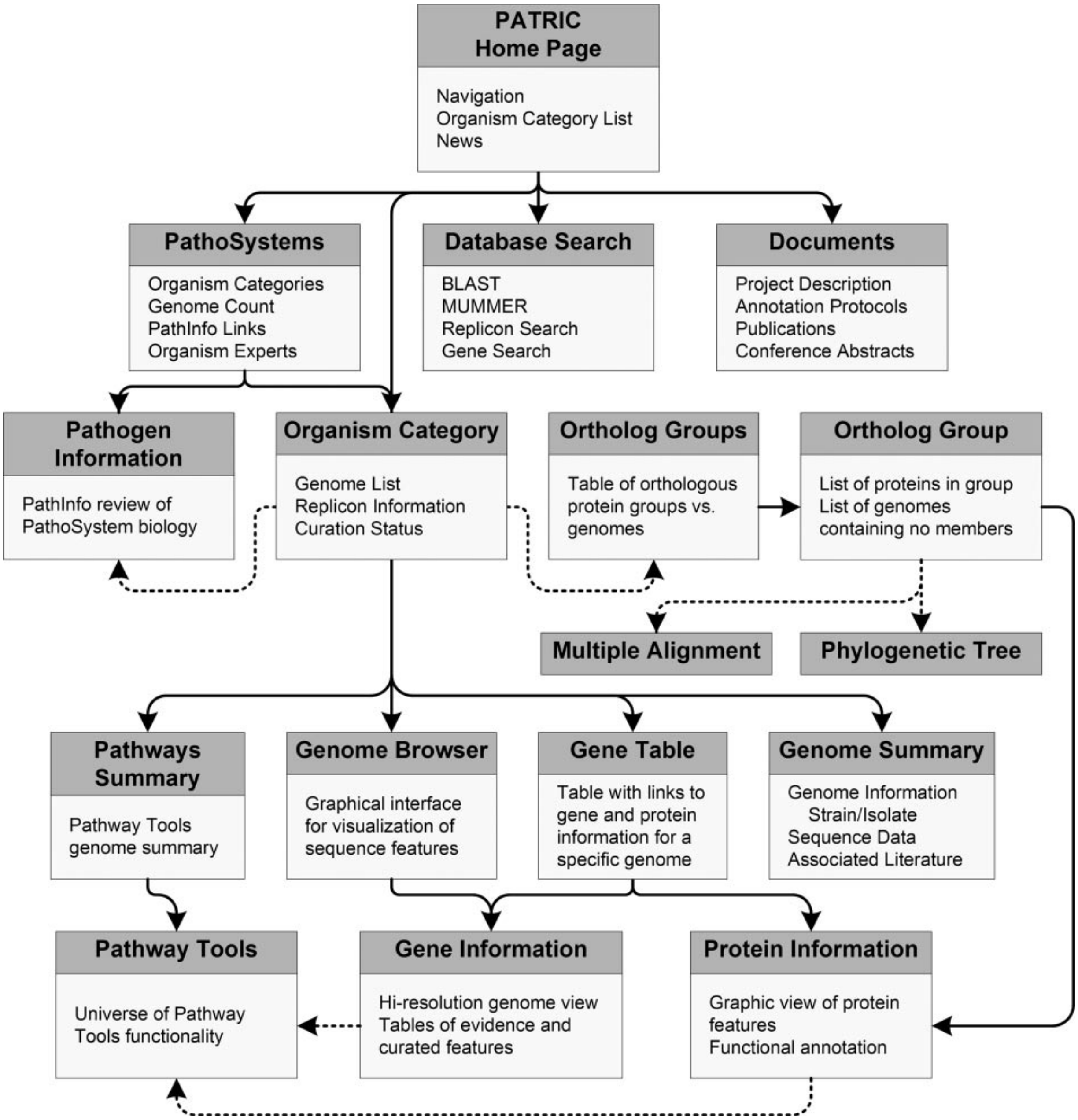

Figure 1. Conceptual map of PATRIC website. Arrows show the relationship between the principal datatype on a page and related data on neighboring pages. Solid arrows represent 'drilling down' to more specific information (e.g. from genome to gene). Dashed arrows represent links between different views of conceptually similar data (e.g. between ortholog group and phylogenetic tree). This figure represents only a subset of the pages and links on the actual website.

of gene order from homologous regions of related genomes and confirmed or rejected by the curation staff. The output is a database with integrated web server that allows users to browse and query the organism's metabolic pathways. This system has been integrated with the PATRIC web site, allowing users to access pathway information for all bacterial reference genomes. The current analysis was based on preexisting RefSeq or GenBank annotations; later releases will incorporate data curated in house, unifying the genomic and pathway versions of the data. The analysis of pathways can facilitate the identification of metabolic choke points, critical enzymes that could be targeted by drugs that may have valuable antimicrobial properties. Pathway analysis can also yield clues to pathogenesis by comparing virulent and avirulent strains and examining the roles of genes not present in both strains.

Website

The PATRIC website is hosted on a Sun Microsystems v20z server running SuSE Linux using the Apache web server. Applications are written in PHP and Perl, accessing data from an Oracle 9i server hosted on a Sun Microsystems E15000 running Sun OS.

The conceptual organization of the website is described in Figure 1. The website's home page contains news, a navigation bar and the list of PATRIC organisms. Users can select their organism of interest from the list to access the corresponding organism category page. This page contains a table of genomes currently in our database with links to the three principal representations of individual genomes: the genome summary, genome browser and gene table. These pages allow users to view a summary of genome sequencing information and to identify specific genes and link to their corresponding gene, protein and pathway information pages. The gene information page displays the output of sequence analysis software run by the annotation pipeline, as well as curated data. Similarly, the protein information page displays InterProScan and TIGRfam alignments and associated information such as GO terms and EC numbers. For bacterial genomes, the pathway information page illustrates the protein's position in the organism's metabolic network and links to a wealth of information provided by PathwayTools.

The organism category page also contains links to a pathogen summary, ortholog group table and a phylogenetic tree based on 16S rRNAs for bacteria or a selected protein family for viruses. For bacterial genomes, detailed pathosystem information is available, provided by the VBI PathInfo documents (14). The ortholog group table shows the presence or absence of reference gene list proteins for each organism in the organism category and provides links to an MSA and tree viewer and the Base-By-Base MSA editor (38) for every ortholog group. Base-By-Base allows users to add sequences to the MSA, recalculate it using Clustal (39), T-Coffee (40) or MUSCLE and generate the corresponding tree using neighbor-joining or a number of clustering algorithms.

The PATRIC website also supports analytical and query tools. A database search page allows user-supplied sequences to be BLASTed against reference and curated sequences from PATRIC organisms. The page also supports MUMMER (41) comparisons between genomes in the database or with a user-supplied sequence. A query tool is available throughout the site by which users can retrieve genes by name, ID, description, as well as GO and EC identifiers and descriptions. Questions, comments and suggestions concerning the website and its contents may be submitted via the 'feedback' page, accessible from the menu bar.

AVAILABILITY

The PATRIC database is hosted at the Virginia Bioinformatics Institute at Virginia Tech and can be accessed via web browser at https://patric.vbi.vt.edu. Sequences and annotation in GFF3 format (see http://song.sourceforge.net/gff3.shtml) can be downloaded by following the 'downloads' link on the main menu bar. GFF3 files are also available through BRC-Central at: http://brc-central.org.

FUTURE DIRECTIONS

This paper presents the first detailed description of the PATRIC website. Future development will advance on several fronts. Genome and proteome curation will continue, complemented by improved tools for query, analysis and visualization. For viruses, we will transition to the more widely accepted ICTV taxonomy (42). The website's user interface is being enhanced to integrate organism-, tool/task- and data-centric approaches to data access, allowing users more efficient and effective access to PATRIC resources. This will be followed up by prioritized curation targeted at potential drug and vaccine targets, virulence factors and genes with differential representation or polymorphisms associated with clinically significant phenotypes. Leveraging another NIAID-funded VBI project, the Administrative Resource for Biodefense Proteomics Research (http://www.proteomicsresource.org/), we plan to integrate expression profiling and proteomics data from pathogen and host to better understand the pathosystem's biology and help the community identify targets for counter-measures. The integration of these disparate data types into a single, easy-to-use system is a goal that we anticipate will enable pathogen researchers to make full use of available data to develop diagnostics, vaccines and therapeutics.

ACKNOWLEDGEMENTS

We would like to thank Chris Upton for making the application Base-By-Base (38) available for incorporation into our website and to Peter Karp for making a similar contribution with his Pathway Tools software (15). This work is funded through NIAID contract HHSN266200400035C to Bruno Sobral. Funding to pay the Open Access publication charges for this article was provided by NIAID contract HHSN266200400035C to Bruno Sobral.

Conflict of interest statement. None declared.

REFERENCES

1. Fauci,A.S. (2003) Biodefence on the research agenda. *Nature*, **421**, 787.
2. Jernigan,D.B., Raghunathan,P.L., Bell,B.P., Brechner,R., Bresnitz,E.A., Butler,J.C., Cetron,M., Cohen,M., Doyle,T., Fischer,M. *et al.* (2002) Investigation of bioterrorism-related anthrax, United States, 2001: epidemiologic findings. *Emerg. Infect. Dis.*, **8**, 1019–1028.
3. Morens,D.M., Folkers,G.K. and Fauci,A.S. (2004) The challenge of emerging and re-emerging infectious diseases. *Nature*, **430**, 242–249.
4. Peterson,J.D., Umayam,L.A., Dickinson,T., Hickey,E.K. and White,O. (2001) The comprehensive microbial resource. *Nucleic Acids Res.*, **29**, 123–125.
5. Chaudhuri,R.R. and Pallen,M.J. (2006) xBASE, a collection of online databases for bacterial comparative genomics. *Nucleic Acids Res.*, **34**, D335–D337.
6. Esteban,D.J., Da Silva,M. and Upton,C. (2005) New bioinformatics tools for viral genome analyses at Viral Bioinformatics—Canada. *Pharmacogenomics*, **6**, 271–280.
7. Kulkarni-Kale,U., Bhosle,S., Manjari,G.S. and Kolaskar,A.S. (2004) VirGen: a comprehensive viral genome resource. *Nucleic Acids Res.*, **32**, D289–D292.
8. Bossi,P., Tegnell,A., Baka,A., Van Loock,F., Hendriks,J., Werner,A., Maidhof,H. and Gouvras,G. (2004) Bichat guidelines for the clinical management of brucellosis and bioterrorism-related brucellosis. *Euro Surveill*, **9**, E15–E16.
9. Kelly,D.J., Richards,A.L., Temenak,J., Strickman,D. and Dasch,G.A. (2002) The past and present threat of rickettsial diseases to military medicine and international public health. *Clin. Infect. Dis.*, **34**, S145–S169.
10. Azad,A.F. and Beard,C.B. (1998) Rickettsial pathogens and their arthropod vectors. *Emerg. Infect. Dis.*, **4**, 179–186.

11. Walker,D.H. (2003) Principles of the malicious use of infectious agents to create terror: reasons for concern for organisms of the genus Rickettsia. *Ann. N Y Acad Sci.*, **990**, 739–742.
12. Centers for Disease Control and Prevention (CDC) (2003) Outbreak of severe acute respiratory syndrome—worldwide, 2003. *MMWR Morb. Mortal. Wkly Rep.*, **52**, 226–228.
13. Fouchier,R.A., Kuiken,T., Schutten,M., van Amerongen,G., van Doornum,G.J., van den Hoogen,B.G., Peiris,M., Lim,W., Stohr,K. and Osterhaus,A.D. (2003) Aetiology: Koch's postulates fulfilled for SARS virus. *Nature*, **423**, 240.
14. He,Y., Vines,R.R., Wattam,A.R., Abramochkin,G.V., Dickerman,A.W., Eckart,J.D. and Sobral,B.W. (2005) PIML: the Pathogen Information Markup Language. *Bioinformatics*, **21**, 116–121.
15. Karp,P.D., Paley,S. and Romero,P. (2002) The Pathway Tools software. *Bioinformatics*, **18** (Suppl. 1), S225–S232.
16. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2006) GenBank. *Nucleic Acids Res.*, **34**, D16–D20.
17. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
18. Ouzounis,C.A. and Karp,P.D. (2002) The past, present and future of genome-wide re-annotation. *Genome Biol.*, **3**, COMMENT2001.
19. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
20. Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.
21. Lukashin,A.V. and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, **26**, 1107–1115.
22. Borodovsky,M. and McIninch,J.D. (1993) GenMark: parallel gene recognition for both DNA strands. *Comput. Chem.*, **17**, 123–133.
23. Suzek,B.E., Ermolaeva,M.D., Schreiber,M. and Salzberg,S.L. (2001) A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics*, **17**, 1123–1130.
24. Tech,M., Pfeifer,N., Morgenstern,B. and Meinicke,P. (2005) TICO: a tool for improving predictions of prokaryotic translation initiation sites. *Bioinformatics*, **21**, 3568–3569.
25. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
26. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
27. Wuyts,J., Perriere,G. and Van De Peer,Y. (2004) The European ribosomal RNA database. *Nucleic Acids Res.*, **32**, D101–D103.
28. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D'Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V., Muller,K.M. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.
29. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038–3049.
30. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
31. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
32. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
33. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
34. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
35. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
36. Retief,J.D. (2000) Phylogenetic analysis using PHYLIP. *Methods Mol. Biol.*, **132**, 243–258.
37. Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.
38. Brodie,R., Smith,A.J., Roper,R.L., Tcherepanov,V. and Upton,C. (2004) Base-By-Base: single nucleotide-level analysis of whole viral genome alignments. *BMC Bioinformatics*, **5**, 96.
39. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
40. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
41. Kurtz,S., Phillippy,A., Delcher,A.L., Smoot,M., Shumway,M., Antonescu,C. and Salzberg,S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.
42. Fauquet,C.M. and Fargette,D. (2005) International Committee on Taxonomy of Viruses and the 3,142 unassigned species. *Virol. J.*, **2**, 64.